DOCUMENT RESUME

ED 481 818 TM 035 329

AUTHOR	Swaminathan, Hariharan; Hambleton, Ronald K.; Sireci, Stephen G.; Xing, Dehui; Rizavi, Saba M.
TITLE	Small Sample Estimation in Dichotomous Item Response Models: Effect of Priors Based on Judgmental Information on the Accuracy of Item Parameter Estimates. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-CTR-98-06
PUB DATE	2003-09-00
NOTE	28p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Estimation (Mathematics); *Item Response Theory; *Sample Size; Sampling
IDENTIFIERS	*Accuracy; Dichotomous Responses; *Item Parameters

ABSTRACT

The primary objective of this study was to investigate how incorporating prior information improves estimation of item parameters in two small samples. The factors that were investigated were sample size and the type of prior information. To investigate the accuracy with which item parameters in the Law School Admission Test (LSAT) are estimated, the item parameter estimates were compared with known item parameter values. By randomly drawing small samples of varying sizes from the population of test takers, the relationship between sample size and the accuracy with which item parameters are estimated was studied. Data used were from the Reading Comprehension subtest of the LAST. Results indicate that the incorporation of ratings of item difficulty provided by subject matter specialists/test developers produced estimates of item difficulty statistics that were more accurate than that obtained without using such information. The improvement was observed for all item response models, including the model used in the LSAT. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 481 818

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. VASELECK

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM035329

LSAC RESEARCH REPORT SERIES

TM TM TR 0246

Small Sample Estimation in Dichotomous Item Response Models: Effect of Priors Based on Judgmental Information on the Accuracy of Item Parameter Estimates

Hariharan Swaminathan
Ronald K. Hambleton
Stephen G. Sireci
Dehui Xing
Saba M. Rizavi

University of Massachusetts Amherst

Law School Admission Council
Computerized Testing Report 98-06
September 2003

LSAC

A Publication of the Law School Admission Council

BEST COPY AVAILABLE

LSAC RESEARCH REPORT SERIES

Small Sample Estimation in Dichotomous Item Response Models: Effect of Priors Based on Judgmental Information on the Accuracy of Item Parameter Estimates

Hariharan Swaminathan
Ronald K. Hambleton
Stephen G. Sireci
Dehui Xing
Saba M. Rizavi

University of Massachusetts Amherst

Law School Admission Council
Computerized Testing Report 98-06
September 2003

A Publication of the Law School Admission Council

Table of Contents

Executive Summary 1

Introduction 2

Item Response Models 2

Estimation of Parameters 3
- *Estimation of Ability Parameters* 3
- *Estimation of Item Parameters* 4
- *Joint Maximum Likelihood Estimation* 4
- *Marginal Maximum Likelihood Estimation* 5

Design of Study 6
- *Sample Size* 6
- *Prior Information* 7
- *Evaluation of the Accuracy of Estimation* 8

Results 9
- *Results for the One-Parameter Model* 9
- *Results for the Two-Parameter Model* 12
- *Results for the Three-Parameter Model* 16

Conclusions 22

References 23

Executive Summary

It is well established that the efficiency of testing can be considerably increased if test takers are administered items that match their ability or proficiency levels. In this adaptive testing scheme, items are administered to test takers sequentially, one at a time or in sets. The item or set of items administered is usually chosen in such a way that it provides maximum information at the proficiency level of the test taker. The feasibility and advisability of computerized adaptive testing is currently being studied by the Law School Admission Council (LSAC).

For adaptive testing to be successful, it is important that a large pool of items be available with items whose item characteristics are known. The recent experiences of testing programs have clearly demonstrated that, without a large item pool, test security can be seriously compromised. One way to maintain a large pool of items is to replenish the pool by administering pretest items to a group of test takers taking an existing test and calculating the statistics for the items. However, administering new items to a large group of test takers increases the exposure rate of these items, compromising test security. One obvious solution is to administer a set of pretest items to a randomly selected small group of test takers. Unfortunately, this solution raises a serious problem: estimating the necessary item-level statistics using small samples of test takers.

Typically in computerized adaptive testing, a mathematical model called item response theory (IRT) is used to describe the characteristics of the test items and the ability level of the test takers. The item-level statistics of this model are commonly referred to as item parameters. In general, large samples are needed to estimate parameters. An issue that needs to be addressed is that of estimating these item parameters using a small sample of test takers. Several research studies have shown that, by incorporating prior information about item parameters, not only can item parameters be estimated more accurately, but estimation can be carried out with smaller sample sizes. The purposes of the current investigation are (i) to examine how prior information about item characteristics can be specified, and (ii) to investigate the relationship between sample size and the specification of prior information on the accuracy with which item parameters are estimated.

The best a priori source for information regarding the difficulty of items in a test is content specialists and test developers. A judgmental procedure for eliciting this information was developed for this study. Once this prior information was obtained, it was combined with data obtained from test takers and the item parameters were estimated.

Since the primary objective of this study was to investigate how incorporating prior information improves estimation of item parameters in small samples, the factors that were investigated were sample size and type of prior information. These two factors were examined with respect to the accuracy with which item parameters were estimated. In order to investigate the accuracy with which item parameters in the Law School Admission Test (LSAT) are estimated, the item parameter estimates were compared with the known item parameter values. By randomly drawing small samples of varying sizes from the population of test takers, the relationship between sample size and the accuracy with which item parameters are estimated was studied. Data from the Reading Comprehension section of the LSAT was utilized.

The results indicate that the incorporation of ratings of item difficulty provided by subject matter specialists/test developers produced estimates of item difficulty statistics that were more accurate than that obtained without using such information. The improvement was observed for all item response models, the evaluated, including the model that is currently used for the LSAT.

This study has demonstrated that using judgmental information about the difficulty of test items can produce dramatic improvements in the estimation of item parameters. This improvement may be sufficient to warrant the routine use of judgmental information in item parameter estimation. However, obtaining judgmental information is time-consuming and costly. The question that arises naturally is whether using some other form of prior information can result in savings and lead to estimates equally as accurate as those obtained by using judgmental information. Several other forms of prior information were used in this study to examine this issue. While using judgmental information produced the most accurate estimates, differences between those estimates obtained using judgmental information and other forms of prior information were not substantial. In order to determine if differences that result from using different forms of prior information are substantial, the effects of using various forms of prior information for item calibration on the routing procedure in an adaptive testing scheme and the estimation of test taker ability need to be investigated. Only through such a study can the improvements offered by incorporating judgmental data as demonstrated in this study and other forms of prior information be fully understood.

BEST COPY AVAILABLE

Introduction

Item response theory (IRT) provides the accepted framework for addressing the fundamental problems in testing: determining the proficiency level of test takers for certification and other reasons, assembly of test items, equating of tests, and examining the potential bias test items may exhibit toward minority or focal groups. In order to fully realize the advantages that item response theory offers, the parameters of item response models must be accurately estimated. These parameters are the ability or proficiency level parameter of a test taker, and the parameters that characterize the item. While estimation of the test taker ability parameter is the ultimate goal of testing, this goal cannot be achieved without determining the parameters that characterize the items. Once the item parameters are determined, items can be "banked," and from this bank, items can be drawn and administered to test takers.

It is well established that the efficiency of testing can be considerably increased if test takers are administered items that match their ability or proficiency levels (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). In this adaptive testing scheme, items are administered to test takers sequentially, one at a time, or in sets. The item or set of items administered is usually chosen in such a way that it provides maximum information at the ability level of the test taker.

For adaptive testing to be successful, it is important that a large pool of items be available with items whose item parameters are known; that is estimated or calibrated using a sample of test takers. The recent experiences of testing programs have clearly demonstrated that without a large item pool, test security can be seriously compromised. One way to maintain a large pool of items is to replenish the pool by administering experimental items to a group of test takers taking an existing test and calibrating the items. However, administering new or experimental items to a large group of test takers increases the exposure rate of these items, compromising test security. One obvious solution is to administer a set of experimental items to a randomly selected small group of test takers. Unfortunately, this solution raises another serious problem: that of estimating item parameters using small samples.

The issue of sample size and its effect on item parameter estimation has been well studied (e.g., Swaminathan & Gifford, 1983). In general, large sample sizes are needed to estimate parameters, particularly in the two- and three-parameter item response models. The issue that needs to be addressed is that of estimating or calibrating items using a small sample of test takers. Swaminathan and Gifford (1982, 1985, 1986) and Mislevy (1986) have shown that, by incorporating prior information about item parameters, not only can item parameters be estimated more accurately, but the estimation can be carried out with smaller sample sizes. The purposes of the current investigation are (i) to examine how prior information can be specified, and (ii) to investigate the relationship between sample size and the specification of prior information on the accuracy with which item parameters are estimated.

This report consists of a brief review of item response models and the issues that surround the estimation of item parameters. The procedure for incorporating prior information is described. The design of the study for investigating the relationship between sample size and prior information is described. The results of the study are presented and the implications for estimating parameters are discussed.

Item Response Models

Dichotomous item response models are classified as one-, two-, or three-parameter models. For all these models, the probability of response u, ($u = 1$ for a correct response and 0 otherwise) to an item, given the item parameters and the ability level of the test taker, is specified by a cumulative probability function, $F(.)$. The common forms of F are the normal and the logistic cumulative probability functions.

In the one-parameter model, the parameter that characterizes the item is called the item difficulty parameter, b. For a test taker with ability θ, the probability of a correct response is 0.5 at $\theta = b_i$. The one-parameter model was developed by Rasch (1960), and hence is commonly referred to as the Rasch model. The probability of a correct response to item i in the Rasch item response model is

$$P(u_i = 1|b_i, \theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}. \quad (1)$$

BEST COPY AVAILABLE

We would like to thank Peter Pashley and Lynda Reese for their assistance in conducting this study.

The probability of a correct response for the two-parameter logistic model is conventionally written in the form

$$P(u_i = 1|a_i, b_i, \theta) = \frac{\exp a_i(\theta - b_i)}{1 + \exp a_i(\theta - b_i)}, \quad (2)$$

where a_i, the discrimination parameter, is the slope of the item response curve at the point of inflection. Whereas in the Rasch model, the log-odds ratios of Rasch item response curves define parallel lines, in the two-parameter logistic models, the lines defined by the log-odds ratios are parallel only when the discrimination parameters are equal.

Motivated by the work of Finney (1952), Birnbaum (1968) introduced the three-parameter logistic item response model given by the item response function

$$P(u_i = 1|c_i, a_i, b_i, \theta) = c_i + (1 - c_i)\frac{\exp a_i(\theta - b_i)}{1 + \exp a_i(\theta - b_i)}. \quad (3)$$

Here the lower asymptote $0 \le c_i < 1$ reflects the probability with which test takers with very low ability or θ values respond correctly to the item i. The parameter c_i is known as the *pseudo chance-level* parameter, or simply as the *guessing parameter*. Empirical studies have shown that with multiple choice items, the three-parameter model fits the item response data better than the Rasch or two-parameter model.

The item response models described above assume that a single dimension θ underlies the test takers' responses to a set of items. The assumption of unidimensionality is an issue of some concern in the measurement literature. While multidimensional item response models have been formulated, the estimation problems associated with multidimensional models are far from being solved. Hence, only the estimation issues concerning unidimensional item response models are discussed in this study.

Estimation of Parameters

Estimation of Ability Parameters

The parameter of ultimate importance in educational testing is the test taker's ability or proficiency level θ. If the item parameters are known a priori, the estimation of θ is straightforward. Let $U = [u_1, u_2, \ldots u_n]$ denote the (n x 1) vector of responses of a test taker to n items. In order to express the joint distribution of U in a tractable form, the assumption of conditional independence, or local independence, has to be made. Assuming that the complete latent space is specified, that is, the number of dimensions that underlie the responses of the population of test takers to a set of items is correctly specified, it can be shown (Anderson, 1959; Lord & Novick, 1968) that the responses of a test taker to n items, conditional on ability, are independent, that is,

$$P(u_1, u_2, \ldots, u_n|\theta, \xi) = \prod_{i=1}^{n} P(u_i|\theta, \xi), \quad (4)$$

where θ is the (r x 1) vector of abilities, and ξ is the vector of item parameters. When it is assumed that $r = 1$, equation 4 holds for unidimensional item response models. Thus, the likelihood function of the observed item responses for a test taker given the item parameters, and, consequently, the maximum of the likelihood function are immediately obtained. The maximum likelihood (ML) estimator of θ can be shown to possess the usual properties of ML estimators with increasing test length (Birnbaum, 1968).

Within a Bayesian framework, if the prior density of θ is $g(\theta|\tau)$ where τ is the vector of known parameters, then the posterior marginal density of θ, $P(\theta \mid U, \tau)$, contains all the information about the parameter θ. The posterior mode or the mean may be taken as a point estimate of θ. When τ is not known, the hierarchical procedure suggested by Lindley and Smith (1972) may be applied. Swaminathan and Gifford (1982, 1985, 1986) applied a two-stage procedure to obtain the joint posterior density of the abilities of N test takers. They assumed that in the first stage

$$\theta_i \sim N(\mu, \phi). \quad (5)$$

BEST COPY AVAILABLE

In the second stage, they assumed that μ was uniform and ϕ had the inverse chi-square density with parameters v and λ. The parameters μ, and ϕ were integrated out of the joint posterior density. The joint modes of the joint posterior density were taken as point estimates of the abilities of the test takers. The joint posterior modes, being weighted estimates of the individual's estimate and the mean of the group, provide more stable estimates of the ability parameters than the mode or the mean of the one-stage Bayes procedure. Because of the complex form of the joint density, Swaminathan and Gifford did not obtain the marginal density or the joint means of the joint posterior density. However, in theory, it is possible to obtain the moments of the posterior density using the approximations suggested by Tierney and Kadane (1986).

An alternative procedure was provided by Bock and Mislevy (1982), who used the mean of the posterior distribution of θ rather than the mode. This expected a posteriori (EAP) was obtained using a single stage procedure, assuming a priori that θ had the standard normal distribution; that is, with mean zero and unit standard deviation.

Estimation of Item Parameters

While the estimation of ability parameters with known item parameters is relatively straightforward, the item parameters must be known or estimated from a calibration sample. If the ability parameters are known, then the item response model becomes a special case of quantal response models, and the estimation of item parameters is again straightforward. However, in general, neither the item parameters nor the ability parameters are known beforehand.

Joint Maximum Likelihood Estimation

The joint estimation of item and ability parameters was proposed by Lord (1953) and Birnbaum (1968). The joint likelihood function of the item and ability parameters, when responses of N test takers on n items are observed, is given by the expression

$$L\left(U_1, U_2, \ldots, U_j, \ldots, U_N \mid \theta, \xi\right) = \prod_{i=1}^{N} \prod_{j=1}^{n} P\left(u_i \mid \theta, \xi\right), \tag{6}$$

where $U_j = [u_{j1}, u_{j2}, \ldots, u_{jn}]$ is the vector of responses of test taker j on n items. It is assumed that the complete latent space is unidimensional, that is, local independence holds.

An examination of the item response models given in Equations 1–3 reveals that the parameters α (a parameter), β (b parameter), and θ are not identified. Linear transformations leave the item response functions invariant, and hence the metric of θ (or β) must be fixed. For convenience, the mean and standard deviation of θ (or β) are usually set at 0 and 1, respectively. In the Rasch model, only the mean of θ (or β) needs to be fixed. Once the metric of θ is fixed, starting with provisional values of θ, the item parameters are estimated by the conventional probit or logit analysis. The item parameters are held fixed at these values, and the values of θ re-estimated. This process is repeated until convergence.

The joint maximum likelihood estimation of item and ability parameters suffers from a major drawback. The ability parameters are *incidental* parameters while the item parameters are *structural* parameters. Neyman and Scott (1948) have shown that the ML estimates of the structural parameters are not consistent in the presence of incidental parameters. While consistent ML estimators of item parameters are not available in the presence of unknown ability parameters for a finite number of items, Haberman (1977) showed that consistent estimates of the Rasch item parameters are obtained as the number of items and the number of test takers increase without limit. Similar results are not available for the two- and three-parameter logistic models. Nevertheless, Swaminathan and Gifford (1983) demonstrated empirically through a series of simulation studies that the estimates of item parameters in the three-parameter model are consistent when the number of items and the number of test takers increase without bound. This empirical finding, although not totally satisfactory, provides some justification for using joint ML estimation with large numbers of items.

Neyman and Scott (1948) also showed that if a minimal sufficient statistic is available for the incidental parameters, conditional maximum likelihood estimators can be devised for the structural parameters. These conditional maximum likelihood estimators enjoy the usual properties of maximum likelihood estimators. A minimal sufficient statistic for the ability parameter is available only for the Rasch model. The total score, r, obtained by summing the item scores, is a minimal sufficient statistic for the ability parameter in the Rasch model. By conditioning on r, Andersen (1970) obtained conditional maximum likelihood estimates of the item parameters. This procedure requires the computation of certain symmetric functions and becomes computationally tedious when the number of items is large.

BEST COPY AVAILABLE

Marginal Maximum Likelihood Estimation

Since a minimal sufficient statistic for the ability parameter is not available for the two- and three-parameter logistic item response models, the conditional maximum likelihood procedure is not applicable for these models. Bock and Lieberman (1970) proposed the marginal maximum likelihood procedure to overcome the difficulties inherent in the joint maximum likelihood procedure. Whereas the joint ML procedure corresponds to the fixed-effects case, the marginal ML procedure corresponds to a mixed model in that the test takers are assumed to be a sample from a known population. The marginal likelihood function is

$$L(U_1, U_2, \ldots, U_i, \ldots, U_N | \xi) = \int \prod_{i=1}^{N} \prod_{j=1}^{n} P(u_i | \theta, \xi)\, g(\theta | \tau)\, d\theta, \tag{7}$$

where $g(\theta \mid \tau)$ is the density function of θ. Bock and Lieberman (1970) took the standard normal density function for $g(\theta \mid \tau)$ and employed Gaussian quadrature to approximate the integral in Equation 7. They solved the resulting likelihood equations using Fisher's method of scoring. In Bock and Lieberman's procedure, the evaluation of the information matrix requires summing over 2^n response patterns and not just the patterns realized in the sample. This made the procedure unwieldy and applicable only to a small number of items. Bock and Aitkin (1981) realized that by fixing certain terms which are functions of item parameters in the likelihood equations at the current values of the parameter estimates, the procedure could be simplified considerably and computational efficiency increased. They pointed out that the fixing of these terms at current values of item parameter estimates could be justified in terms of the EM algorithm of Dempster, Laird, and Rubin (1977). Bock and Aitkin (1981), however, noted that their algorithm is not strictly the same as the general EM algorithm. For random variables in the models not belonging to the exponential family, Dempster, Laird, and Rubin (1977) take the expected value of the logarithm of the likelihood function while Bock and Aitkin take the expected value of the likelihood function. It should be pointed out that the Bock and Aitkin application of the EM algorithm was not the first application of this algorithm to item response models. Sanathanan and Blumenthal (1978) applied the EM algorithm to the Rasch model to estimate the parameters τ of $g(\theta \mid \tau)$. Their procedure, however, is restricted to the Rasch model and does not generalize to other item response models.

Rigdon and Tsutakawa (1983) and Tsutakawa (1984) applied an extended form of the EM algorithm appropriate when the random variables in the models do not belong to the exponential family. They applied the procedure developed by Dempster, Rubin, and Tsutakawa (1981) for estimating linear effects in mixed models to obtain marginal maximum likelihood estimates of item parameters in the one- and two-parameter item response models. They also provided simplified computational procedures for estimating the item parameters, the ability parameters, and the variance of the ability distribution.

Bayesian procedures. While the marginal maximum likelihood procedures have theoretical advantages over the joint maximum likelihood procedures, the estimates of the discrimination parameter α and the chance-level parameter γ pose considerable problems in that these parameters are often poorly estimated and the estimates frequently drift off into inadmissible regions. Bayesian procedures show considerable promise in terms of their ability to successfully address these issues.

Bayesian procedures for estimating item parameters were proposed by Swaminathan and Gifford, who, in a series of papers (Swaminathan & Gifford, 1982, 1985, 1986), provided a hierarchical procedure for the one-, two-, and three- parameter models based on the Lindley-Smith approach (Lindley & Smith, 1972). They assumed that the item difficulty parameters and the ability parameters are exchangeable and obtained the joint density of the item and ability parameters, marginalized with respect to the parameters of the ability and item difficulty distributions, that is,

$$p(\xi, \theta | U, \delta, \zeta) = L(U | \theta, \xi) \int \int p(\theta | \tau)\, p(\xi | \eta)\, p(\eta | \zeta)\, p(\tau | \delta)\, d\tau\, d\eta, \tag{8}$$

where U contains the responses of N test takers on n items. In particular, Swaminathan and Gifford assumed that the parameters β_j were independently and identically normally distributed with mean μ and variance ϕ; the parameter α_j had a chi-density with parameters v and ω; and the parameter γ had a beta density with parameters p and q. They also provided procedures for specifying the parameters of the prior distributions. Swaminathan and Gifford obtained joint modal estimates of the posterior distribution using the Newton-Raphson procedure to solve the modal equations. Their results were promising in that the drift of the parameter estimates was arrested and the parameters were estimated more accurately than the joint ML procedure.

BEST COPY AVAILABLE

A problem with the Bayesian approach of Swaminathan and Gifford was that it was not free from the criticisms that faced joint estimation of the parameters. Another problem was that different forms of prior distributions had to be specified for the various item parameters given the varying nature of the item parameters. One solution to this problem is to specify a multi-parameter density for the priors such as a multi-parameter beta distribution for the item parameters.

Mislevy (1986), Tsutakawa (1992), and Tsutakawa and Lin (1986) have provided a marginalized Bayes modal estimation procedure by integrating out the ability parameters and using the EM algorithm to estimate the parameters. Mislevy (1986) has suggested transforming the discrimination and chance-level parameters, so that a multivariate normal prior for the item parameters can be specified. The specification of multivariate normal priors for the item parameters removes the problem inherent in the separate prior specifications proposed by Swaminathan and Gifford. However, Bayes modal estimates are not invariant with respect to transformations and hence the Bayes modal estimates of the transformed parameters cannot be transformed back to the original metric of the parameters. Nevertheless, the marginalized Bayes procedure of Mislevy is an improvement over the joint procedure of Swaminathan and Gifford. The marginalized Bayes procedure is currently implemented in the BILOG program (Mislevy & Bock, 1990) for estimating item parameters in the dichotomous case, albeit with separate forms for the priors—a normal prior for β, a log-normal prior for α, and a beta prior for γ. The procedure suggested by Tsutakawa (1992) and Tsutakawa and Lin (1986) for specifying priors is basically different from that suggested by Swaminathan and Gifford and Mislevy. Tsutakawa and Lin (1986) suggested an ordered bivariate beta distribution for the item response function at two ability levels, while Tsutakawa (1992) suggested the ordered Dirichlet prior on the entire item response function. These approaches are promising, but no extensive research has been done to date comparing this approach with other Bayesian approaches.

Design of the Study

The primary objective of this study was to investigate estimation of item parameters in small samples and to determine the specification of prior information that will result in accurate estimation in small samples. Given this, the factors that were investigated were sample size and type of prior information. These two factors were examined with respect to the accuracy with which item parameters were estimated in the one-, two-, and three-parameter item response models.

In order to investigate the accuracy with which item parameters are estimated, it is necessary to compare the item parameter estimates with the "true" item parameter values. Typically, such an investigation is carried out using simulated data since true values of item and ability parameters cannot be known a priori. With simulated data, general conditions can be simulated. One drawback, however, is that the item parameter values selected for the study may not conform to real testing situations. More importantly, the distributions of ability and item parameters may conform too closely to the prior distributions when Bayesian procedures are investigated, possibly limiting the generalizability of the results to real data.

Fortunately, the estimation procedures can be investigated with real data—in this study, Law School Admission Test (LSAT) data from the Law School Admission Council (LSAC). Since the test was administered to a large group of test takers, calibrating the items with the entire population of test takers will yield true item parameters. With small samples randomly drawn from the population of test takers, varying the sample size and estimating the item parameters will yield the relationship between sample size and the accuracy with which item parameters are estimated. Moreover, the Bayesian procedures will yield untainted information regarding the effects of prior specifications on the accuracy of estimation.

Parameter estimation in the three-item response models were investigated in this study. In order to obtain true parameter values for the parameters in the one-, two-, and three-parameter models, each model was fitted to the data for the LSAT Reading Comprehension section. Only the 21 items for which judges provided ratings of difficulty were used. The estimates corresponding to the relevant parameters in each model were taken as the true values.

Sample Size

One of the primary concerns in calibration is the minimal sample size that is needed to provide reasonably accurate estimates of item parameters. Hence, one of the factors that was examined in the study was sample size. Sample size was varied from a relatively small sample (n = 100) to a modest sample size (n = 500). Six levels of sample size were used in this study: 100, 150, 200, 300, 400, and 500. These sample sizes were chosen so that the effect of prior information could be studied carefully in a narrow range of sample size values.

BEST COPY AVAILABLE

Prior Information

As has been demonstrated by several researchers, accurate estimation of item parameters in small samples, particularly in the two- and three-parameter models, can only be accomplished through a Bayesian approach. In order to implement a Bayesian procedure, prior information must be specified on item parameters. Prior information can be specified in a variety of forms.

Previous research on Bayesian estimation employed priors that were, in some sense, arbitrary with simulated data. While this approach provided information regarding the effects of priors that reflected the distribution of true item parameters as well as priors that deviated from the true distribution of the item parameters (Swaminathan & Gifford, 1982, 1985, 1986), they did not, and could not, reflect the information practitioners had regarding the items. In order to study the effect of prior information on the accuracy of estimation, which is based on the knowledge that test developers have regarding the items, a new procedure was developed. This procedure involved extracting information from test developers in an objective manner and transforming this information into a prior distribution which could then be interfaced with a Bayesian procedure (see Hambleton, Sireci, Swaminathan, Xing, & Rizavi, 1999).

Prior information on item difficulty—judgmental information. The procedure for obtaining judgmental information regarding item parameters from a panel of subject matter specialists and test developers involved (i) training subject matter specialists and test developers as to the nature of item parameters; (ii) eliciting information, independently, from them regarding the difficulty levels of items; and (iii) using a consensus building approach, allowing them, if they chose, to revise their initial estimates of difficulty level of the items. The item difficulty information provided by the subject matter specialists and test developers was in the form of the proportion of test takers who, according to the raters' belief, would respond correctly to the item. This information had to be translated to correspond to the Item Response Theory (IRT) item-difficulty parameter, and a prior distribution specified to enable the information to be interfaced with the Bayesian procedure.

Prior distribution for item difficulty parameter. The judgmental rating obtained regarding the difficulty level of an item is the proportion of the test takers who respond correctly. The proportion is on the interval [0,1] and must be mapped onto to the scale of IRT item difficulty parameter, that is, mapped onto the real line.

Let p denote the proportion of test takers who, according to a rater, respond correctly to an item. A convenient transformation that carries the proportion-correct score onto the scale of the IRT item difficulty parameter is

$$b_0 = -\Phi^{-1}(p),$$

where Φ is the normal ogive function, that is, p is the area under the normal curve to the left of the normal deviate b_0. The negative sign is to ensure that an item with a high p-value (an easy item) will have a negative value for the IRT item difficulty parameter.

In determining the normal deviate, the following approximation, attributed to L. Tucker (Bock & Jones, 1968), was used to facilitate computing:

$$b_0 = \frac{U\left(a_1 - a_2 U^2 + a_3 U^4\right)}{\left(1 - a_4 U^2 + a_5 U^4\right)},$$

where

$$U = p - \tfrac{1}{2},\ a_1 = 2.5101,\ a_2 = 12.2043,\ a_3 = 11.2502,\ a_4 = 5.8742,\ \text{and},\ a_5 = 7.9587.$$

The prior distribution for the item difficulty parameter was taken as the normal density function with mean equal to the average of the raters' transformed p-values. Three values were used as the standard deviation (SD) of the distribution.

(1) A standard deviation of one, reflecting a "tight" prior.

(2) A standard deviation of two, reflecting a diffuse prior.

(3) The standard deviation corresponding to the standard deviation of the judges' transformed ratings.

BEST COPY AVAILABLE

It should be pointed out that the standard deviation of the judges' transformed ratings may not reflect the standard deviation of the prior distribution. This is because the judges provided only what they thought was the "difficulty" level of the item; they did not indicate how "confident" they were with the rating they provided. For example, all the judges may provide the same value for p. This will result in a standard deviation of zero for the prior distribution. This, however, does not reflect the confidence the raters had about their ratings. Despite this, the standard deviation of the transformed p-values was taken as one of the measures of standard deviation for the purpose of investigation.

In addition to the three prior distributions based on judgmental data described above, six other prior distributions were considered. These are

1. normal prior with mean equal to the transformed true p-value and standard deviation, one;
2. normal prior with mean equal to the transformed true p-value and standard deviation, two;
3. normal prior with mean equal to the transformed sample p-value and standard deviation, one;
4. normal prior with mean zero and standard deviation, one; and
5. normal prior with mean zero and standard deviation, two;

In addition, a condition using no prior information was included.

The prior specifications described under (1) and (2) are critical in that they establish the veracity of the premise underlying the study. The premise underlying the study is that if subject matter specialists can provide information regarding the difficulty level of the item, this information can be used as the prior information for the difficulty parameters. If the premise is true, then, clearly, the true p-value provides the most accurate prior information for the difficulty parameters, and, hence, using this value as the mean of the prior distribution should result in the most accurate estimation of the difficulty parameters. If the estimation using the true p-value to set the prior produces poor results, then it can be argued that asking subject matter specialists to provide information regarding the difficulty level of the item will not be useful.

It can be argued along the same lines that if information regarding item difficulty is useful, a less costly method of obtaining this information is by computing the sample p-value rather than by assembling a panel of experts. Hence, the accuracy of estimation obtained by using the sample p-value needs to be investigated.

One disadvantage of using the sample p-value is that in small samples, the p-value is relatively unstable. An alternate approach is to ignore the information available in the sample and specify a prior that is sample independent. Normal priors with mean zero (standard deviations of one and two) were used to compare the estimation accuracies obtained with sample-based and sample-free priors.

The accuracy of estimation that may result from using a Bayesian approach must be compared with the classical statistical approach where no priors are used. Hence, in the last condition, no priors were specified.

The subject matter specialists were not asked to provide information regarding either the discrimination or the lower asymptote parameters. This was because no intuitive approach by which the experts could be asked to provide information regarding these parameters was available. Hence, sample-free priors were used for the discrimination and lower asymptote parameters.

The prior distribution for the discrimination parameter a in the two- and the three-parameter models was taken as the log-normal distribution, that is, it was assumed that the natural logarithm of a was distributed normally with mean zero and standard deviation one (Mislevy, 1986). The prior distribution for the lower asymptote parameter, c, in the three-parameter model was taken as a beta distribution (Swaminathan & Gifford, 1986) with a mean of 0.2 (corresponding to a test taker choosing one of the five options in a multiple choice item randomly), and a standard deviation of .0095 (corresponding to weight of 20 observations attached to the mean).

The item parameters were estimated using the program BILOG (Mislevy & Bock, 1990) Version 7.1.

Evaluation of the Accuracy of Estimation

In order to evaluate the accuracy with which the item parameters were estimated, the estimates were compared to the true values based on a sample of 5,000 test takers. Since the evaluation of the accuracy of estimation cannot be assessed without carrying out replications, 100 replications were carried out for each condition.

BEST COPY AVAILABLE

The accuracy with which item parameters are estimated can be ascertained by computing the discrepancy between the estimate and the true value. Let τ_i be the true value of an item parameter (a, b, or c) for item i, and t_{ki} its estimate in the kth replication. The Mean Square Error (MSE_i), for item i for an item parameter (a, b, or c) is defined as

$$MSE_i = \frac{\sum_{k=1}^{R}(t_{ki} - \tau_i)^2}{R}.$$

Gifford and Swaminathan (1990) have shown that when replications are carried out, the Mean Square Error defined above can be decomposed into Squared Bias and Variance, defined as

$$Squared\ Bias_i = (\bar{t}_i - \tau)^2$$

and

$$Variance_i = \frac{\sum_{k=1}^{R}(t_{ik} - \bar{t}_i)^2}{k}.$$

Thus, MSE = Squared Bias + Variance. This decomposition provides an explanation of the value obtained for the MSE. A large MSE could result from either bias in the estimation or a large sampling fluctuation. For example, Bayesian procedures generally result in estimates which have a larger bias and smaller sampling variance than maximum likelihood estimates. These quantities can be averaged over items to provide summary indices. For descriptive purposes, the square roots of these quantities averaged over items are reported: root mean square error, ($RMSE$), Bias, and standard error (SE).

Since there were six sample sizes, ten priors (including no prior), and three item response models, there were 180 conditions to be replicated. In all, 18,000 computer runs were executed. In order to extract the information from the BILOG output and to provide summary information such as $RMSE$, Bias, and SE, a computer program was written to interface with BILOG.

Results

The results of the study are presented for the parameters of the one-, two-, and the three-parameter models. Graphical displays are also provided for $RMSE$ and Bias for the estimation of parameters in these models.

Results for the One-parameter Model

Table 1 contains the average root mean square values for the difficulty parameter for the ten prior specifications and the six sample sizes. For all sample sizes, the Bayesian procedure resulted in improved estimation when compared to the "no prior" or marginal maximum likelihood (MML) procedure. The only exception to this trend resulted with the prior based on the judges' ratings, which used the standard deviation of the transformed ratings. This is not a surprising result, given the reasons provided earlier. As expected, the prior based on the true p-value yielded the most accurate estimates. The priors based on judges' ratings with sample independent standard deviations for the priors yielded results identical to the priors based on the true p-values.

BEST COPY AVAILABLE

TABLE 1
Average root mean square error of item difficulty parameter estimates for the one-parameter model under various prior distributions on difficulty

Prior Distribution		Sample Size					
		100	150	200	300	400	500
No prior		0.3932	0.3222	0.2800	0.2354	0.1995	0.1793
Normal priors:							
Mean	*SD*						
Transformed "true" *p*-value	2	0.3900	0.3204	0.2788	0.2347	0.1990	0.1790
Transformed "true" *p*-value	1	0.3807	0.3151	0.2753	0.2327	0.1977	0.1781
Mean transformed judges' ratings	2	0.3911	0.3211	0.2793	0.2323	0.1992	0.1792
Mean transformed judges' ratings	1	0.3860	0.3185	0.2776	0.2337	0.1986	0.1788
	SD of transformed ratings						
Mean transformed judges' ratings	2	0.4930	0.4225	0.3762	0.2311	0.2656	0.2425
Transformed sample *p*-value	1	0.3928	0.3220	0.2799	0.2353	0.1994	0.1793
Transformed sample *p*-value	2	0.3918	0.3214	0.2795	0.2351	0.1992	0.1792
0	1	0.3928	0.3220	0.2799	0.2353	0.1994	0.1793
0		0.3919	0.3150	0.2795	0.2351	0.1993	0.1792

In order of accuracy of estimation, the procedures are prior based on true value with *SD* 1.0, prior based on judges transformed *p*-values with *SD* 1.0, prior based on true value with *SD* 2.0, prior based on judges transformed *p*-values with *SD* 2.0, prior based on sample *p*-values with *SD* 1.0 and prior based on the normal distribution with mean zero and *SD* 1, prior based on sample *p*-values with *SD* 2.0 and prior based on the normal distribution with mean zero and *SD* 2, no prior, and finally, prior based on judges' transformed *p*-values with the standard deviation based on the observed standard deviation. This trend was evident at all sample sizes. However, all the procedures, with the exception of the procedure based on the observed judges' standard deviation, produced indistinguishable results when the sample sizes were 400 and 500. These results are displayed graphically in Figure 1.

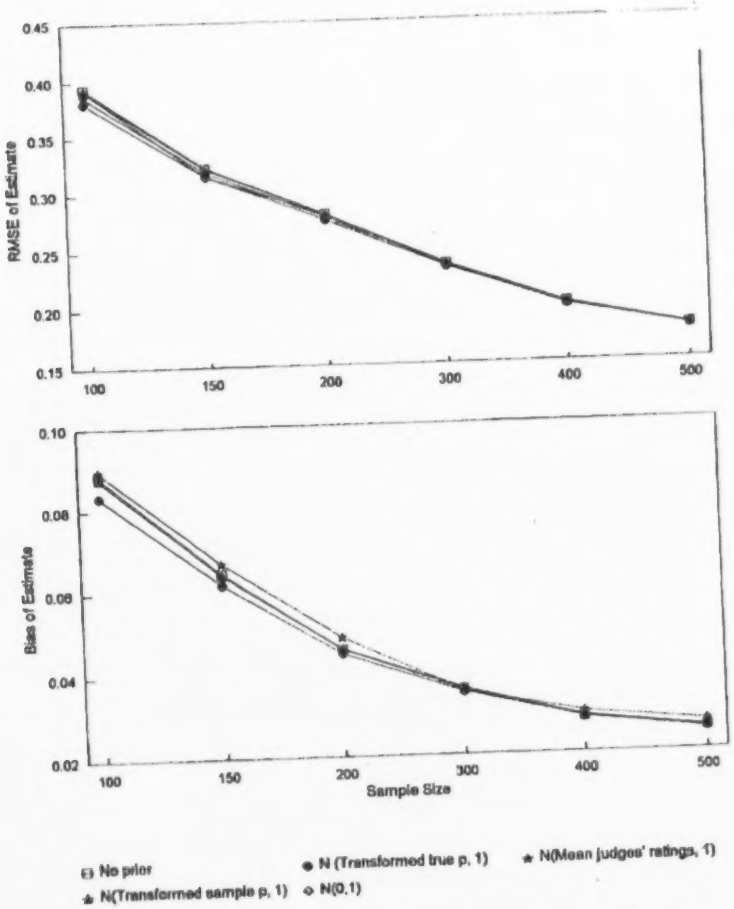

FIGURE 1. *Effect of prior on difficulty on estimation of difficulty under the one-parameter model as a function of sample size.*

The results corresponding to bias in estimation are provided in Table 2. Surprisingly, most of the Bayesian procedures showed less bias than the estimates based on no prior on difficulty. The exceptions to this were those judges' ratings with a tight prior, that is, *SD* of 1.0. The prior based on the *SD* of the judges' ratings showed the most bias. Given that this procedure produces unacceptable results in all situations, this procedure will be henceforth omitted from the discussions of results. As sample sizes increase, the differences among the procedures diminish. A graphic display of the bias results for the one-parameter model is provided in Figure 1.

TABLE 2
Average bias of item difficulty parameter estimates for the one-parameter model under various prior distributions on difficulty

Prior Distribution		Sample Size 100	150	200	300	400	500
No prior		0.0882	0.0643	0.0459	0.0356	0.0286	0.0254
Normal priors:							
Mean	*SD*						
Transformed "true" *p*-value	2	0.0869	0.0636	0.0456	0.0355	0.0285	0.0253
Transformed "true" *p*-value	1	0.0832	0.0617	0.0448	0.0351	0.0283	0.0251
Mean transformed judges' ratings	2	0.0875	0.0643	0.0462	0.0352	0.0287	0.0256
Mean transformed judges' ratings	1	0.0895	0.0668	0.0488	0.0353	0.0297	0.0269
	SD of transformed ratings						
Mean transformed judges' ratings	2	0.4068	0.3413	0.2983	0.2350	0.1961	0.1793
Transformed sample *p*-value	1	0.0880	0.0639	0.0459	0.0357	0.0285	0.0254
Transformed sample *p*-value	2	0.0876	0.0641	0.0460	0.0360	0.0283	0.0255
0	1	0.0880	0.0642	0.0459	0.0356	0.0287	0.0254
0		0.0878	0.0640	0.0460	0.0356	0.0287	0.0254

In general, the incorporation of priors resulted in very modest improvements in the estimation of the difficulty parameter in the one-parameter model. Ignoring the prior based on true *p*-values, the judges' ratings yielded the most accurate estimates. Since the trend lines across samples do not cross, the improvement in estimation obtained using prior information cannot be converted to savings in terms of sample size.

Results for the Two-parameter Model

The results pertaining to the accuracy of estimation for the difficulty parameter in the two-parameter model are given in Table 3. The results follow the pattern that was exhibited for the one-parameter model. The procedures, in order of their accuracy, from most accurate to least accurate, as determined by *RMSE*, are prior based on true *p*-value with *SD* 1.0, prior based on judges' transformed *p*-values with *SD* 1.0, prior based on true *p*-value with *SD* 2.0, prior based on judges' transformed *p*-values with *SD* 2.0, prior based on sample *p*-value with *SD* 1.0 and prior based on the normal distribution with mean zero and *SD* 1, prior based on sample *p*-value with *SD* 2.0 and prior based on the normal distribution with mean zero and *SD* 2, no prior, and finally, prior based on judges' transformed *p*-values with the observed standard deviation. This trend persisted across all sample sizes, with the differences in the *RMSE* across procedures diminishing with increasing sample size. It is clear that the difficulty parameter is less well estimated in the two-parameter model than in the one-parameter model. This is to be expected as the introduction of more parameters in the model decreases the accuracy with which the parameters are estimated in small samples. Figure 2 provides a visual display of these results. (Note: The procedure with the observed standard deviation for the judges' ratings is omitted since inclusion of this distorts the scale.)

BEST COPY AVAILABLE

TABLE 3
Average root mean square error of item difficulty parameter estimates for the two-parameter model under various prior distributions on difficulty

Prior Distribution		Sample Size 100	150	200	300	400	500
No prior		0.4119	0.3445	0.3072	0.2565	0.2249	0.1998
Normal priors:							
Mean	*SD* 2	0.4093	0.3426	0.3058	0.2557	0.2243	0.1993
Transformed "true" *p*-value	1	0.4020	0.3374	0.3017	0.2535	0.2226	0.1978
Transformed "true" *p*-value	2	0.4101	0.3431	0.3061	0.2558	0.2244	0.1994
Mean transformed judges' ratings	1	0.4051	0.3395	0.3031	0.2539	0.2229	0.1983
Mean transformed judges' ratings	*SD* of transformed ratings	0.4172	0.3551	0.3194	0.2691	0.2352	0.2156
Transformed sample *p*-value	2	0.4111	0.3437	0.3065	0.2561	0.2246	0.1995
Transformed sample *p*-value	1	0.4091	0.3416	0.3046	0.2552	0.2237	0.1987
0	2	0.4110	0.3436	0.3064	0.2561	0.2246	0.1995
0	1	0.4089	0.3413	0.3042	0.2549	0.2236	0.1986

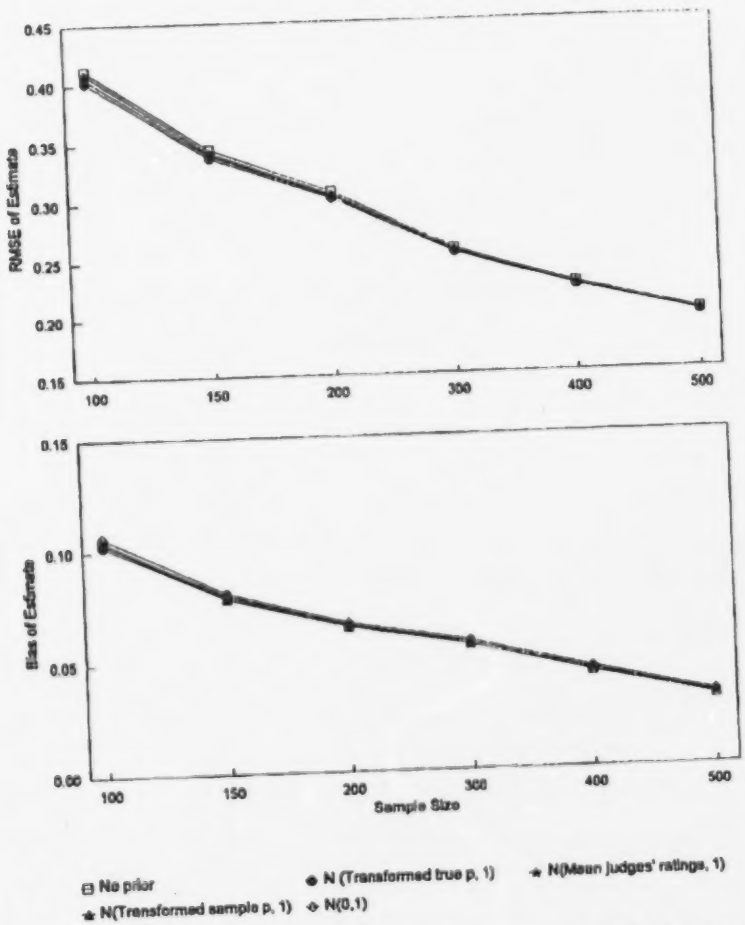

FIGURE 2. *Effect of prior on difficulty on estimation of difficulty under the two-parameter model as a function of sample size.*

BEST COPY AVAILABLE

Table 4 contains the results pertaining to bias in the estimation of the difficulty parameter. Comparison of priors with standard deviation of one and two reveal the predictable result: tighter priors result in more bias than diffuse priors; the only exception being the procedure that was based on true *p*-value. In this case, a tighter prior yielded less biased estimates compared with the corresponding diffuse prior. Surprisingly, the procedure with no prior on the difficulty parameter did not yield the least bias. Although bias is present, the differences among the procedures are negligible, a positive result where bias is concerned. Figure 2 provides a visual display of these findings. (Note: The procedure with the observed standard deviation for the judges' ratings is omitted.)

TABLE 4
Average bias of item difficulty parameter estimates for the two-parameter model under various prior distributions on difficulty

		Sample Size					
Prior Distribution		100	150	200	300	400	500
No prior		0.1041	0.0783	0.0650	0.0563	0.0430	0.0314
Normal priors:							
Mean	*SD*						
Transformed "true" *p*-value	2	0.1037	0.0782	0.0651	0.0565	0.0432	0.0317
Transformed "true" *p*-value	1	0.1026	0.0783	0.0656	0.0572	0.0441	0.0323
Mean transformed judges' ratings	2	0.1037	0.0785	0.0652	0.0562	0.0430	0.0316
Mean transformed judges' ratings	1	0.1039	0.0797	0.0663	0.0561	0.0435	0.0324
	SD of transformed						
Mean transformed judges' ratings	ratings	0.2699	0.2257	0.2002	0.1604	0.1342	0.1258
Transformed sample *p*-value	2	0.1041	0.0785	0.0650	0.0566	0.0432	0.0317
Transformed sample *p*-value	1	0.1044	0.0791	0.0659	0.0573	0.0441	0.0324
0	2	0.1046	0.0788	0.0654	0.0567	0.0434	0.0318
0	1	0.1062	0.0806	0.0668	0.0579	0.0449	0.0330

Table 5 contains the *RMSE* values for the estimation of the discrimination parameter. It should be noted that different priors were placed only on the difficulty parameter. The same prior distribution was imposed on the discrimination parameter in all the procedures. Given this, a comparison of the no prior condition with the prior conditions reveals that placing a prior on the difficulty had a positive effect on the estimation of the discrimination parameter. Tighter priors on the difficulty parameter, as determined by the standard deviation of the prior distribution, yielded more accurate estimation. The only exceptional result that sets the estimation of the difficulty parameter apart from the estimation of the discrimination parameter is the prior distribution with the observed standard deviation of the judges' ratings. This prior distribution resulted in the most accurate estimation of the discrimination parameter. Apart from this, the smallest *RMSE* was observed for the standard normal prior, followed by the prior based on judges' ratings with a standard deviation of one.

TABLE 5
Average root mean square error of item discrimination parameter estimates for the two-parameter model under various prior distributions on difficulty

		Sample Size					
Prior Distribution		100	150	200	300	400	500
No prior		0.1135	0.0918	0.0791	0.0685	0.0587	0.0554
Normal priors:							
Mean	*SD*						
Transformed "true" *p*-value	2	0.1117	0.0906	0.0783	0.0680	0.0583	0.0551
Transformed "true" *p*-value	1	0.1067	0.0873	0.0762	0.0665	0.0572	0.0542
Mean transformed judges' ratings	2	0.1114	0.0903	0.0781	0.0678	0.0582	0.0550
Mean transformed judges' ratings	1	0.1055	0.0863	0.0757	0.0661	0.0569	0.0539
	SD of transformed						
Mean transformed judges' ratings	ratings	0.0969	0.0865	0.0825	0.0720	0.0638	0.0598
Transformed sample *p*-value	2	0.1119	0.0907	0.0784	0.0680	0.0583	0.0551
Transformed sample *p*-value	1	0.1078	0.0877	0.0765	0.0667	0.0572	0.0542
0	2	0.1103	0.0897	0.0778	0.0676	0.0580	0.0549
0	1	0.1021	0.0843	0.0745	0.0654	0.0563	0.0535

These results with respect to *RMSE* are graphically displayed in Figure 3. An examination of the figure reveals that (using linear interpolation), without prior information, a sample size of 125 is needed to achieve the same degree of accuracy as that obtained with a prior based on judges' ratings with a standard deviation of one. This translates into a saving of 25% in terms of sample size, at a sample size value of 150. It should be noted that the savings, in terms of sample size for the difficulty parameter, is considerably less.

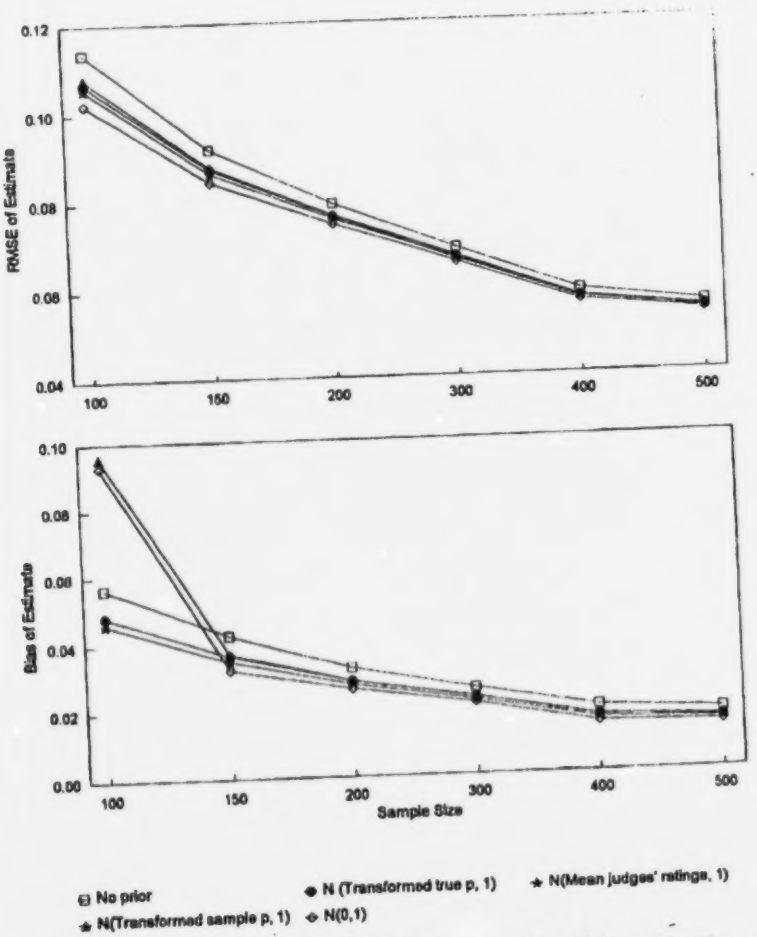

FIGURE 3. *Effect of prior on difficulty on estimation of discrimination under the two-parameter model as a function of sample size.*

An examination of Table 6 reveals that the prior based on judges' ratings with a standard deviation of one produced the least biased estimates. The normal priors with mean zero and standard deviations of one and two and sample *p*-value based priors resulted in the most biased estimates. The results in Table 6 and Figure 2 show that as sample size increases, the bias decreases rapidly. The prior based on the judges' observed standard deviation produced estimates with the largest bias while the standard normal prior resulted in the smallest bias.

BEST COPY AVAILABLE

TABLE 6
Average bias of item discrimination parameter estimates for the two-parameter model under various prior distributions on difficulty

Prior Distribution		Sample Size 100	150	200	300	400	500
No prior		0.0564	0.0424	0.0321	0.0256	0.0194	0.0177
Normal priors:							
Mean	*SD*						
Transformed "true" *p*-value	2	0.0542	0.0408	0.0310	0.0248	0.0187	0.0171
Transformed "true" *p*-value	1	0.0482	0.0364	0.0282	0.0228	0.0166	0.0154
Mean transformed judges' ratings	2	0.0536	0.0403	0.0307	0.0246	0.0185	0.0169
Mean transformed judges' ratings	1	0.0460	0.0347	0.0273	0.0221	0.0159	0.0149
	SD of transformed						
Mean transformed judges' ratings	ratings	0.0634	0.0604	0.0598	0.0502	0.0437	0.0394
Transformed sample *p*-value	2	0.0978	0.0409	0.0311	0.0249	0.0187	0.0171
Transformed sample *p*-value	1	0.0959	0.0368	0.0284	0.0229	0.0166	0.0155
0	2	0.0971	0.0395	0.0302	0.0243	0.0181	0.0167
0	1	0.0930	0.0321	0.0260	0.0211	0.0147	0.0140

Results for the Three-Parameter Model

The entries in Table 7 refer to the accuracy of estimation of the difficulty parameter in the three-parameter model. The *RMSE* values are larger than in the one- and the two-parameter models indicating that the difficulty parameter is less well estimated in the three-parameter model than in the other two models when the sample size is small (500 or less). A comparison of the procedures reveals that the prior based on the judges' ratings with a standard deviation of one produced the most accurate estimates across all sample sizes. The prior based on the true *p*-value produced the second most accurate estimates. In general, a tighter prior produced more accurate estimates than the corresponding diffuse prior. This trend was evident across all sample sizes, with the *RMSE* decreasing steadily as the sample size increases. A graphical display of the accuracy of estimation is provided in Figure 4. A comparison of the *RMSE* with the prior based on judges' ratings with the estimates obtained with no prior reveals that a sample size of 150 with no prior yields the same level of accuracy as that obtained with a sample size of 100 when using a prior based on the judges' ratings a saving of 50% in terms of sample size. This saving decreases as the sample increases.

TABLE 7
Average root mean square error of item difficulty parameter estimates for the three-parameter model under various prior distributions on difficulty

Prior Distribution		Sample Size 100	150	200	300	400	500
No prior		0.5182	0.4745	0.4309	0.3771	0.3455	0.3399
Normal priors							
Mean	*SD*						
Transformed "true" *p*-value	2	0.4985	0.4609	0.4188	0.3724	0.3422	0.3357
Transformed "true" *p*-value	1	0.4775	0.4421	0.4037	0.3652	0.3363	0.3278
Mean transformed judges' ratings	2	0.4962	0.4583	0.4166	0.3702	0.3404	0.3335
Mean transformed judges' ratings	1	0.4735	0.4371	0.3985	0.3595	0.3315	0.3219
	SD of transformed						
Mean transformed judges' ratings	ratings	0.5329	0.5051	0.4859	0.4645	0.4502	0.4398
Transformed sample *p*-value	2	0.5018	0.4632	0.4207	0.3737	0.3431	0.3363
Transformed sample *p*-value	1	0.4894	0.4507	0.4104	0.3698	0.3397	0.3302
0	2	0.5023	0.4632	0.4218	0.3744	0.3437	0.3372
0	1	0.4920	0.4526	0.4122	0.3713	0.3410	0.3323

BEST COPY AVAILABLE

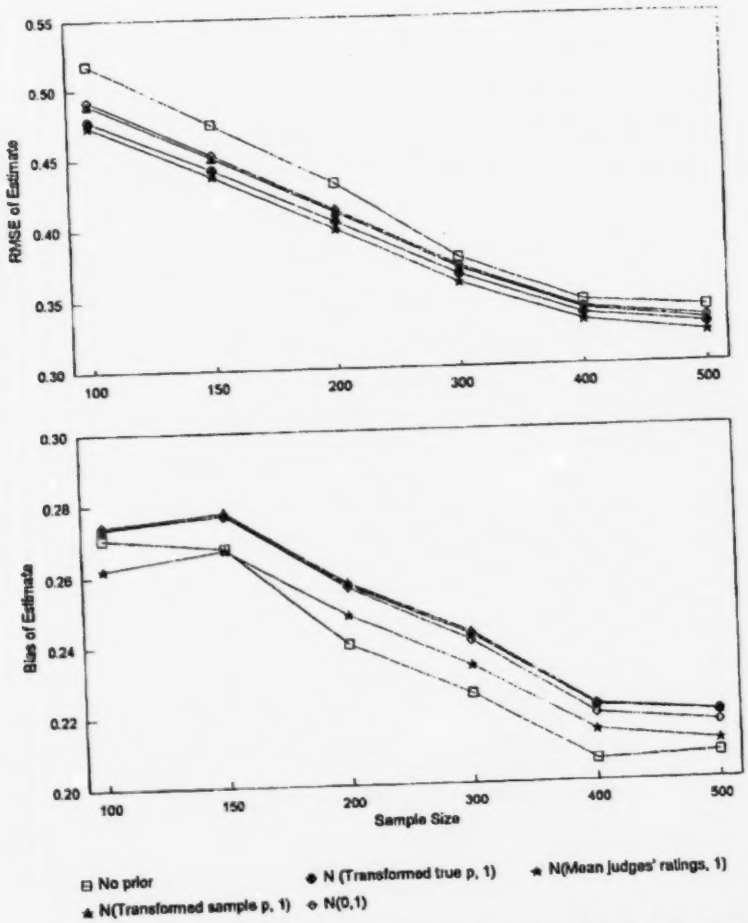

FIGURE 4. *Effect of prior on difficulty on estimation of difficulty under the three-parameter model as a function of sample size.*

The bias in the estimation of the difficulty parameter is presented in Table 8. The prior based on judges' ratings yielded the least biased estimates. Tighter prior yielded less biased estimates than the corresponding diffuse prior in all cases. Surprisingly, estimates with no prior specification were more biased than their counterparts based on prior information. The bias in the estimates, however, does not seem to decrease as rapidly as with the one- and the two-parameter models as sample size increases.

TABLE 8
Average bias of item difficulty parameter estimates for the three-parameter model under various prior distributions on difficulty

Prior Distribution		Sample Size 100	150	200	300	400	500
No prior		0.2705	0.2677	0.2399	0.2254	0.2061	0.2074
Normal priors							
Mean	*SD*						
Transformed "true" *p*-value	2	0.2670	0.2702	0.2437	0.2292	0.2095	0.2094
Transformed "true" *p*-value	1	0.2734	0.2767	0.2566	0.2421	0.2214	0.2191
Mean transformed judges' ratings	2	0.2633	0.2668	0.2407	0.2260	0.2068	0.2063
Mean transformed judges' ratings	1	0.2620	0.2671	0.2481	0.2333	0.2145	0.2109
	SD of transformed						
Mean transformed judges' ratings	ratings	0.4589	0.4448	0.4336	0.4234	0.4129	0.4056
Transformed sample *p*-value	2	0.2671	0.2706	0.2440	0.2294	0.2097	0.2095
Transformed sample *p*-value	1	0.2735	0.2776	0.2573	0.2428	0.2218	0.2192
0	2	0.2671	0.2701	0.2436	0.2289	0.2090	0.2087
0	1	0.2741	0.2773	0.2560	0.2404	0.2192	0.2161

With respect to the estimation of the discrimination parameter, Table 9 reveals that for a sample size of 100, the standard normal prior produced the most accurate estimates; the prior based on judges' ratings with a standard deviation of one produced a similar *RMSE*. With a sample size of 150 and larger, the prior based on judges' ratings with a standard deviation of one produced more accurate estimates. As with the difficulty parameter, a tighter prior produced the most accurate estimates than the corresponding diffuse prior. A closer examination of Table 9 reveals that using a prior based on judges' ratings with a standard deviation of one results in more than 100% savings in terms of sample size when no prior is used; that is, a sample size of 100 with prior yields the same degree of accuracy as that obtained with a sample size of 200 when no prior is used. Figure 5 provides a graphical display of the results described above.

TABLE 9
Average root mean square error of item discrimination parameter estimates for the three-parameter model under various prior distributions on difficulty

Prior Distribution		Sample Size 100	150	200	300	400	500
No prior		0.2275	0.1979	0.1724	0.1580	0.1472	0.1337
Normal priors							
Mean	*SD*						
Transformed "true" *p*-value	2	0.1958	0.1778	0.1568	0.1477	0.1385	0.1273
Transformed "true" *p*-value	1	0.1726	0.1577	0.1451	0.1360	0.1283	0.1202
Mean transformed judges' ratings	2	0.1921	0.1743	0.1543	0.1455	0.1373	0.1260
Mean transformed judges' ratings	1	0.1700	0.1552	0.1437	0.1339	0.1271	0.1198
	SD of transformed						
Mean transformed judges' ratings	ratings	0.2132	0.2101	0.2113	0.2093	0.2081	0.2101
Transformed sample *p*-value	2	0.1968	0.1785	0.1572	0.1480	0.1387	0.1274
Transformed sample *p*-value	1	0.1748	0.1591	0.1461	0.1368	0.1297	0.1207
0	2	0.1906	0.1735	0.1549	0.1453	0.1367	0.1258
0	1	0.1691	0.1567	0.1461	0.1366	0.1300	0.1222

BEST COPY AVAILABLE

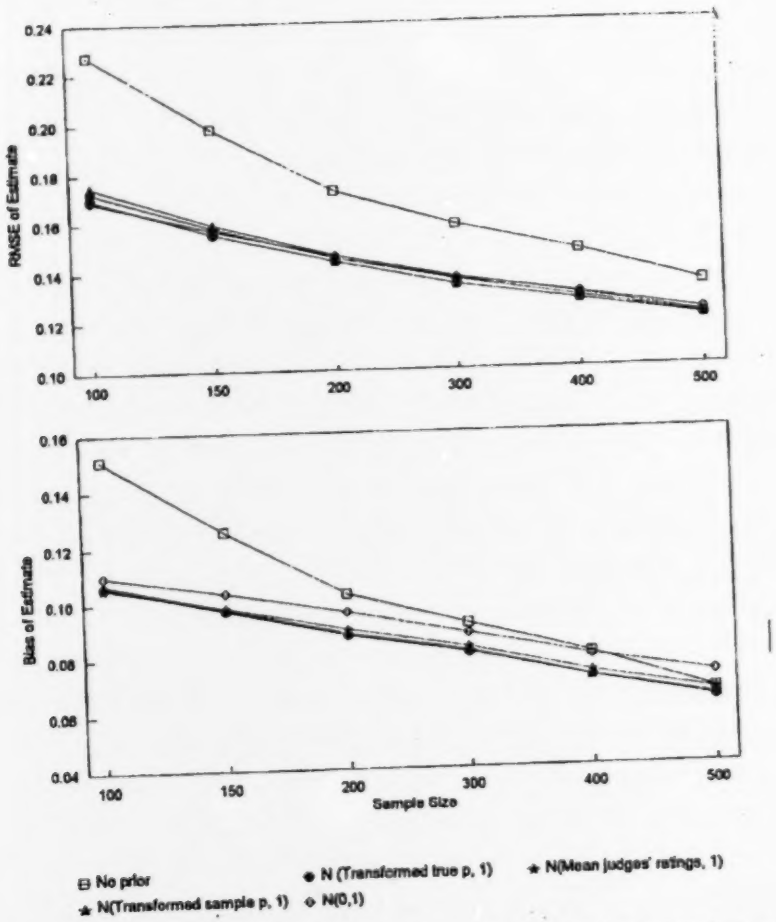

FIGURE 5. *Effect of prior on difficulty on estimation of discrimination under the three-parameter model as a function of sample size.*

Using a prior based on judges' ratings with a standard deviation of one yielded the least biased estimates of the discrimination parameter (Table 10). For a sample size of 100, the estimation procedure that was not based on prior information resulted in estimates that were 50% more biased than those based on priors obtained using judges' ratings. At larger sample values, the procedures did not differ from each other with respect to bias.

TABLE 10
Average bias of item discrimination parameter estimates for the three-parameter model under various prior distributions on difficulty

Prior Distribution		Sample Size 100	150	200	300	400	500
No prior		0.1510	0.1256	0.1026	0.0915	0.0801	0.0666
Normal priors							
Mean	*SD* 2	0.1239	0.1074	0.0898	0.0829	0.0722	0.0610
Transformed "true" *p*-value	1	0.1060	0.0972	0.0878	0.0811	0.0718	0.0636
Transformed "true" *p*-value	2	0.1200	0.1042	0.0876	0.0809	0.0701	0.0595
Mean transformed judges' ratings	1	0.1055	0.0981	0.0900	0.0828	0.0736	0.0662
Mean transformed judges' ratings	*SD* of transformed ratings	0.1917	0.1937	0.1958	0.1986	0.1983	0.2018
Transformed sample *p*-value	2	0.1246	0.1079	0.0901	0.0832	0.0723	0.0612
Transformed sample *p*-value	1	0.1068	0.0978	0.0883	0.0816	0.0718	0.0642
0	2	0.1189	0.1037	0.0874	0.0809	0.0704	0.0598
0	1	0.1097	0.1035	0.0961	0.0880	0.0794	0.0727

The most dramatic improvements in estimation were observed with the estimation of the *c*-parameter (Table 11). The prior based on judges' ratings with a standard deviation of one produced the most accurate estimates. The estimates in order from most accurate to least accurate are prior based on judges transformed *p*-values with *SD* 1.0, prior based on sample *p*-value with *SD* 1.0, prior based on true *p*-value with *SD* 1.0, prior based on the normal distribution with mean zero and *SD* 1, the priors based on true *p*-value with *SD* 2.0, on judges' transformed *p*-values with *SD* 2.0, on sample *p*-values with *SD* 2.0, and on the normal distribution with mean zero and *SD* 2 (all producing equally accurate estimates). The estimate based on no prior resulted in the least accurate estimates (with the exception of the estimate based on judges' ratings with the observed standard deviation).

TABLE 11
Average root mean square error of item lower asymptote parameter estimates for the three-parameter model under various prior distributions on difficulty

Prior Distribution		Sample Size 100	150	200	300	400	500
No prior		0.0707	0.0697	0.0678	0.0664	0.0645	0.0637
Normal priors							
Mean	*SD* 2	0.0673	0.0670	0.0651	0.0643	0.0625	0.0619
Transformed "true" *p*-value	1	0.0644	0.0637	0.0622	0.0615	0.0597	0.0590
Transformed "true" *p*-value	2	0.0664	0.0660	0.0642	0.0635	0.0618	0.0612
Mean transformed judges' ratings	1	0.0625	0.0618	0.0603	0.0597	0.0582	0.0574
Mean transformed judges' ratings	*SD* of transformed ratings	0.0746	0.0754	0.0771	0.0813	0.0839	0.0854
Transformed sample *p*-value	2	0.0672	0.0670	0.0651	0.0643	0.0625	0.0619
Transformed sample *p*-value	1	0.0641	0.0635	0.0620	0.0614	0.0597	0.0589
0	2	0.0672	0.0669	0.0653	0.0645	0.0626	0.0621
0	1	0.0651	0.0646	0.0633	0.0625	0.0607	0.0604

The accuracy results for the *c*-parameter are graphically displayed in Figure 6. The figure demonstrates that the procedure based on judges' ratings with a standard deviation of one as prior yielded more accurate estimates of the *c*-parameter with a sample of 100 than the estimate based on a sample of 500 without prior information specified, a 500% savings in terms of sample size!

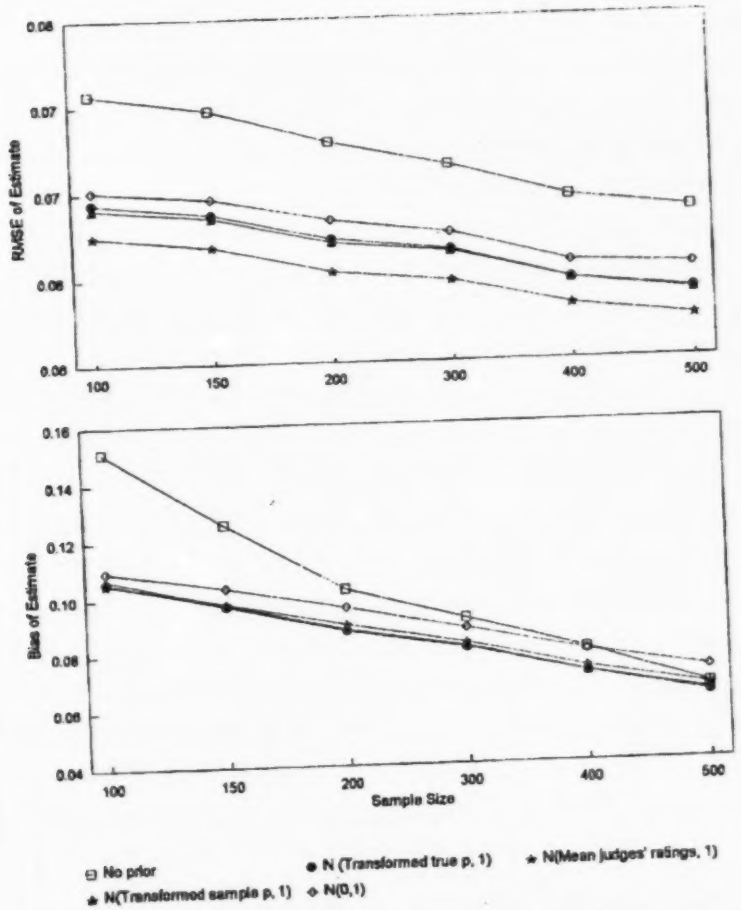

FIGURE 6. *Effect of prior on difficulty on estimation of lower asymptote under the three-parameter model as a function of sample size.*

With respect to bias (Table 12), the procedures which incorporated prior information yielded less biased estimates than the procedure that did not use prior information on the difficulty parameter. The size of the bias compared to the *RMSE* values indicates that the inaccuracy in estimation is primarily due to bias in estimation.

BEST COPY AVAILABLE

TABLE 12
Average bias of item lower asymptote parameter estimates for the three-parameter model under various prior distributions on difficulty

Prior Distribution		Sample Size 100	150	200	300	400	500
Normal priors							
Transformed "true" *p*-value	2	0.0605	0.0591	0.0560	0.0538	0.0510	0.0494
Transformed "true" *p*-value	1	0.0594	0.0578	0.0554	0.0531	0.0502	0.0485
Mean transformed judges' ratings	2	0.0597	0.0583	0.0552	0.0530	0.0503	0.0487
Mean transformed judges' ratings	1	0.0577	0.0560	0.0537	0.0515	0.0488	0.0469
Mean transformed judges' ratings	*SD* of transformed ratings	0.0682	0.0697	0.0722	0.0773	0.0801	0.0821
Transformed sample *p*-value	2	0.0605	0.0591	0.0560	0.0538	0.0510	0.0495
Transformed sample *p*-value	1	0.0594	0.0578	0.0554	0.0532	0.0503	0.0485
0	2	0.0604	0.0591	0.0561	0.0538	0.0510	0.0496
0	1	0.0602	0.0587	0.0565	0.0542	0.0512	0.0498

Conclusions

The purpose of this study was to investigate if estimation of item parameters in item response models can be improved by incorporating prior information, especially in the form of judgements regarding the difficulty level of the items provided by content specialists and test developers. It was anticipated that incorporation of such information will improve estimation in small samples.

The results provided above indicate that the incorporation of ratings provided by subject matter specialists/test developers regarding item difficulty in the form of a prior distribution produced estimates that were more accurate than that obtained without using such information. The improvement was observed for all item response models. The improvement observed for the estimation of the item difficulty parameter in the one-parameter model was modest; that is, although improvements were observed, the gains may not warrant the cost incurred in obtaining judgmental information.

The improvement observed in the estimation of item difficulty and discrimination parameters in the two-parameter model through incorporating judgmental information was somewhat greater than that observed in the one-parameter model. While there was only a modest improvement in the estimation of the difficulty parameter in the two-parameter model, the improvement in the estimation of the discrimination parameter was more substantial. Using prior information in the form of judges' ratings, the estimates obtained with a sample size of 100 yielded the same level of accuracy as that obtained with a sample size of 150 when no prior information was used. This translates into a 50% improvement in the estimation of the discrimination parameter in the two-parameter model.

The improvements obtained with the three-parameter model were clearly substantial. In the estimation of the difficulty parameter, using prior information in the form of judgmental ratings yielded an improvement of 50% over not using prior information when a sample size of 100 was used. In the estimation of the discrimination parameter, an improvement of 100% was observed using judgmental ratings for a sample size of 100; that is, the accuracy obtained with the use of judgmental ratings for a sample size of 100 was comparable to the accuracy level obtained with a sample size of 200 when no prior information was used. The result was most dramatic with the estimation of the *c*-parameter. The accuracy obtained with a sample size of 100 with judgmental ratings was superior to that obtained with a sample size of 500 when no prior information was used. This corresponds to a 500% cost savings in terms of sample size.

The procedure used in this study used judgmental information about item difficulty only. If procedures can be developed for obtaining judgmental information about the discrimination and lower asymptote parameters, considerable improvement in the estimation of item parameters may result. A judgmental procedure for eliciting such information from subject matter specialists and test developers is currently being investigated (see Hambleton, et al., 1999).

This study has demonstrated that in the three-parameter model, using judgmental information about the difficulty parameter produces dramatic improvements in the estimation of the discrimination and lower asymptote parameters. This improvement may be sufficient to warrant the use of judgmental information in item calibration. However, obtaining judgmental information is time-consuming and costly. The question that arises naturally is whether using some other form of prior information can result in savings and lead to estimates equally accurate as those obtained by using judgmental information. Several other forms of prior information were used in this study to examine this issue. While using judgmental information produced the most accurate estimates, difference between those estimates obtained using judgmental information and other forms of prior information, although not substantial, nevertheless exist. In order to determine if

BEST COPY AVAILABLE

differences that result from using different forms of prior information are substantial, the effects of using various forms of prior information for item calibration on the routing procedure in an adaptive testing scheme and the estimation of proficiency need to be investigated. Only through such a study can the improvements offered by incorporating judgmental data as demonstrated in this study and other forms of prior information be fully understood.

References

Anderson, T. W. (1959). Some scaling models and estimation procedures in the latent class model. In O. Grenander (Ed.), *Probability and statistics*, The Harold Cramfer Volume. New York: Wiley.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B*, 32, 283–301.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459.

Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. San Francisco: Holden-Day.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179–197.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4), 431–444.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 76, 341–353.

Finney, D. J. (1952). *Probit analysis: A statistical treatment of the sigmoid response curve*. Cambridge, England: Cambridge University Press.

Gifford, J. A., & Swaminathan, H. (1990). Accuracy, bias, and effect of priors on the Bayesian estimators of parameters in item response models. *Applied Psychological Measurement*, 14(1), 33–43.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, 5, 815–841.

Hambleton, R. K., Sireci, S. G., Swaminathan, H., Xing, D., & Rizavi, S. (1999). *Anchor-based methods for judgmentally estimating item difficulty parameters* (Laboratory of Psychometric and Evaluative Research Report No. 310). Amherst, MA: University of Massachusetts, School of Education.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 1–41.

Lord, F. M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18*, 57–75.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: *Item analysis and test scoring with binary logistic models* [computer program]. Mooresville, IN: Scientific Software.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika, 16,* 1–32.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rigdon, S. E., & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48,* 567–574.

Sanathanan, L., & Blumenthal, S. (1978). The logistic model and estimation of latent structure. *Journal of the American Statistical Association, 73,* 794–799.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7,* 175–191.

Swaminathan, H., & Gifford, J. A. (1983). *Estimation of parameters in the three-parameter latent trait model.* In D. Weiss (Ed.), New horizons in testing (pp. 13–30). New York: Academic Press.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50,* 175–191.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51,* 581–601.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association, 81,* 82–86.

Tsutakawa, R. K. (1984). Estimation of two-parameter logistic item response curves. *Journal of Educational Statistics, 9,* 263–276.

Tsutakawa, R. K. (1992). Prior distributions for item response curves. *British Journal of Mathematical and Statistical Psychology, 45,* 51–74.

Tsutakawa, R. K., & Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika, 51,* 251–267.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").